Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the Earth System Grid Federation eco-system

S. Fiore[1], M. Płóciennik[2], C. Doutriaux[3], C. Palazzo[1], J. Boutte[3], T. Żok[2], D. Elia[1], M. Owsiak[2], A. D'Anca[1], Z. Shaheen[3], R. Bruno[4], M Fargetta[4], M. Caballer[5], G. Moltó[5], I. Blanquer[5], R. Barbera[4,6], M. David[7], G. Donvito[4], D. N. Williams[3], V. Anantharaj[8], D. Salomoni[4], and G. Aloisio[1,9]

[1]Euro-Mediterranean Center on Climate Change Foundation (CMCC), Italy
[2]Poznan Supercomputing and Networking Center (PSNC), Poland
[3]Lawrence Livermore National Laboratory (LLNL), California, USA
[4]Italian National Institute of Nuclear Physics (INFN), Italy
[5]Universitat Politècnica de València (UPV), Spain
[6] University of Catania, Italy
[7]Laboratório de Instrumentação e Fisica Experimental de Particulas (LIP), Portugal
[8]Oak Ridge National Laboratory (ORNL), Tennessee, USA
[9]University of Salento, Italy

***Abstract*—A case study on *climate models intercomparison data analysis* addressing several classes of multi-model experiments is being implemented in the context of the EU H2020 INDIGO-DataCloud project. Such experiments require the availability of large amount of data (multi-terabyte order) related to the output of several climate models simulations as well as the exploitation of scientific data management tools for large-scale data analytics. More specifically, the paper discusses in detail a use case on precipitation trend analysis in terms of requirements, architectural design solution, and infrastructural implementation. The experiment has been tested and validated on CMIP5 datasets, in the context of a large scale distributed testbed across EU and US involving three ESGF sites (LLNL, ORNL, and CMCC) and one central orchestrator site (PSNC).**

Keywords-big analytics, workflow management, cloud computing, ESGF, INDIGO-DataCloud.

I. INTRODUCTION

The increased models resolution in the development of comprehensive Earth System Models is rapidly leading to very large climate simulations output that pose significant scientific data management challenges in terms of data sharing, processing, analysis, visualization, preservation, curation, and archiving [1-3].

In this domain, large scale global experiments for climate model intercomparison (CMIP) have led to the development of the Earth System Grid Federation (ESGF [4-5]), a federated data infrastructure involving a large set of data providers/modelling centers around the globe, which includes the European contribution - regarding the ENES [6] community – through the IS-ENES project.

From an infrastructural standpoint, ESGF provides a production-level support for search & discovery, browsing and access to climate simulation data and observational data products. ESGF has been serving the Coupled Model Intercomparison Project Phase 5 (CMIP5) experiment, providing access to 2.5PB of data for the Intergovernmental Panel on Climate Change (IPCC) [7] Assessment Reports 5 [8], based on consistent metadata catalogues. More precisely, the Coupled Model Intercomparison Project (CMIP) has been established by the Working Group on Coupled Modelling [9] (WGCM) under the World Climate Research Programme [10] (WCRP).

It provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation and data access. This framework enables a diverse community of scientists to analyse General Circulation Models (GCMs) in a systematic fashion, a process that serves to facilitate models improvement.

CMIP5 has promoted a standard set of model simulations in order to:

- evaluate how realistic the models are in simulating the recent past;
- provide projections of future climate change on two time scales, near term (out to about 2035) and long term (out to 2100 and beyond); and
- understand some of the factors responsible for differences in model projections, including quantifying some key feedbacks such as those involving clouds and the carbon cycle.

In such a context, running a multi-model data analysis experiment is very challenging, as it requires the availability of large amount of data (multi-terabyte order) related to multiple climate models simulations as well as scientific data management tools for large-scale data analytics.

The remainder of this work is organized as it follows. Section II provides the current workflow for the multi-model climate data analysis in the CMIP context, whereas Section III presents the paradigm shift needed to address such large-

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

scale data challenges. Section IV introduces the case study on climate change research faced by the INDIGO-DataCloud project [11-12] focusing on the key requirements, general workflow, expected results, and a specific use case on precipitation trend analysis. Section V provides the architectural design solution, whereas Section VI the infrastructural implementation performed in a real testbed across US and Europe (presented in Section VII). Section VIII presented the Kepler user interface, whereas the added value from the proposed INDIGO-DataCloud solution is discussed in Section IX. Section X draws the final conclusions and highlights future work.

II. Current Workflow for Multi-Model Climate Data Analysis in the ESGF/CMIP5 Context

Today, ESGF acts mainly as a large-scale, federated data sharing facility. In such a context to perform a CMIP5-based data analysis, end-users must download all the needed datasets from the ESGF data nodes on their local machines before starting to run the analysis steps. Such a preparatory step represents a strong barrier for climate scientists, as the data download can take (depending on the amount of data needed by the experiment) a significant amount of time. Moreover, analysing large datasets involves running multiple data operators, from widely adopted set of command line tools (mostly sequential CLI). This is usually done via scripts on the client side and requires climate scientists to take care of, implement and replicate workflow-like control logic aspects in their scripts - along with the expected application-level part. Yet, in the current workflow, end-users need to have system management/ICT skills to install and update all the required data analysis tools/libraries on their local machines. Finally, the large volumes of data and the strong I/O requirements pose additional challenges related to performance, which requires a substantial re-design effort at the storage level both in terms of I/O interfaces and physical storage layout to address current issues.

III. Paradigm Shift Joining Big Data, HPC and Cloud Computing

The issues mentioned in the previous section are going to get worse and unmanageable in the exabyte era. To address and overcome them, a strong paradigm shift must be envisioned. In the INDIGO-DataCloud project we explore a different approach based on (i) data-intensive facilities running big analytics frameworks jointly with (ii) server-side analysis capabilities, and (iii) cloud computing solutions.

Data intensive facilities close to the different storage hierarchies are needed to address high-performance scientific data management. On top of them, parallel applications for big data analysis (exploiting MPI, OpenMP, MapReduce paradigms as well as machine learning and data mining techniques) should provide a new generation of "tools" for climate scientists.

Server-side approaches will intrinsically and drastically reduce data movement. Download will only relate to the final results of an analysis (e.g., images, maps, reports and summaries typically megabytes or even kilobytes). Such approach will strongly reduce the amount of data downloaded on the client side as well as the complexity related to the analysis software to be installed on client machines. Moreover, they will also enable managing raw data, intermediate products, final outputs, workflows, lineage information and users sessions on the back-end. Finally, server-side approaches will also require a strong effort on interoperability and standard interfaces in order to build highly interoperable tools and environments for climate data analysis. In this regard, both the Research Data Alliance (RDA [13]) and ESGF are already working on these topics with valuable contributions, among the others, on big data analytics, array-databases, persistent identifiers, and standard interfaces for server-side processing.

In such a landscape joining HPC and big data, *cloud technologies* will help on deploying in a flexible and dynamic manner analytics applications/tools as containers or virtual machines, thus enabling highly scalable and elastic scenarios in both private clouds and cluster environments.

IV. Case Study on Climate Change Research

With specific regard to the CMIP5 context, the case study on climate change research proposed in the INDIGO-DataCloud project focuses on the following classes of multi-model data analysis: anomalies analysis, trend analysis and climate change signal analysis. All of them require the access to a federated data repository (e.g. managed by ESGF data nodes) as well as running complex analytics experiment with tens/hundreds of data analytics operators.

Throughout the paper, the general "environment" of the case study will relate to: (i) *multi-model data analysis inter-comparison challenges*, (ii) *addressed on CMIP5 data*, (iii) *which are made available through the IS-ENES/ESGF infrastructure*.

A. Key Requirements

With regard to the proposed case study, a set of 15 requirements has been gathered during the first months of the INDIGO project [14] after a requirements elicitation process with a group of climate scientists from CMCC, LLNL and ORNL. The most relevant ones are reported in the following.

- *Efficiency/Scalability*: running massive inter-comparison data analysis can be very challenging due to the large volume of the involved datasets (e.g. multi-terabyte order). There is a strong need to provide scalable solutions (e.g. HPC-, HTC-based) and different paradigms (e.g. server-side).
- *Interoperability*: there is a general eco-system for the scientific community that has be taken into account (e.g. existing data repositories, interfaces, security infrastructure, data formats, standards, specifications, tools, etc.). Interoperability with the existing ESGF/IS-ENES infrastructure is key.
- *Workflow support*: data analysis inter-comparison experiments are based on multiple (e.g. tens/hundreds) data operators. Workflow tools could help managing the complexity of these experiments at different levels (multi-site and single-site) and

increase the re-usability of specific workflow templates in the community.

- *Metadata management*: it represents a complementary (w.r.t to "data") aspect that must be taken into consideration both from a technical (e.g. metadata tools) and a scientific (e.g. data semantics) point of view.
- *Easy to use analytics environments*: providing an easy-to-use and integrated analytics environment for climate model inter-comparison could represent an added value to enable scientific research at such large scale. From a technical point of view it also relates to having easy deployment procedures (e.g. cloud-based) to enable a larger adoption by the community.
- *Flexible, elastic and dynamic environments*: it must be considered that the data analysis workload can considerably vary over time (in this regard the CMIP experiments are a very significant example), so proper solutions from an ICT point of view must be investigated.

B. General Workflow

Jointly with the climate scientists involved into the experiment, the main steps related to a general workflow example for our case study have been defined. They are presented in the following sub-sections.

1) Experiment definition: Starting from a user interface (graphical or command-line) the climate scientist should be able to choose/define a specific type of data analysis. In this regard, an associated data analysis workflow could be: (i) either selected from a repository (addressing re-usability), customized and re-used, or (ii) composed on the fly by the user (and then – eventually - stored in the workflows repository for further re-use). It should include the entire workflow description (a detailed tasks and dependencies definition related to the expected data analysis process). Input parameters are provided at this stage.

2) Experiment run: The data analysis workflow should be submitted to the infrastructure. Computational/storage resources should be allocated for the data analysis taking into special account data locality aspects. Access to the data from multiple data centers as well as reduction tasks could be required for multi-model analysis. The data analysis tasks produce intermediate data as well as final products. Workflow solutions would be strongly required to support running these experiments, jointly with tasks monitoring capabilities.

3) Results access, visualization, and publication: The results should be made easily available to the end-user through a dedicated interface for download, visualization and possibly further analysis. It should be also possible to publish the results of a specific analysis on dedicated catalogues. The user interface should provide analytics, exploration and visualization capabilities. To this end, already existing and well-known tools in the community should be integrated in the general eco-system.

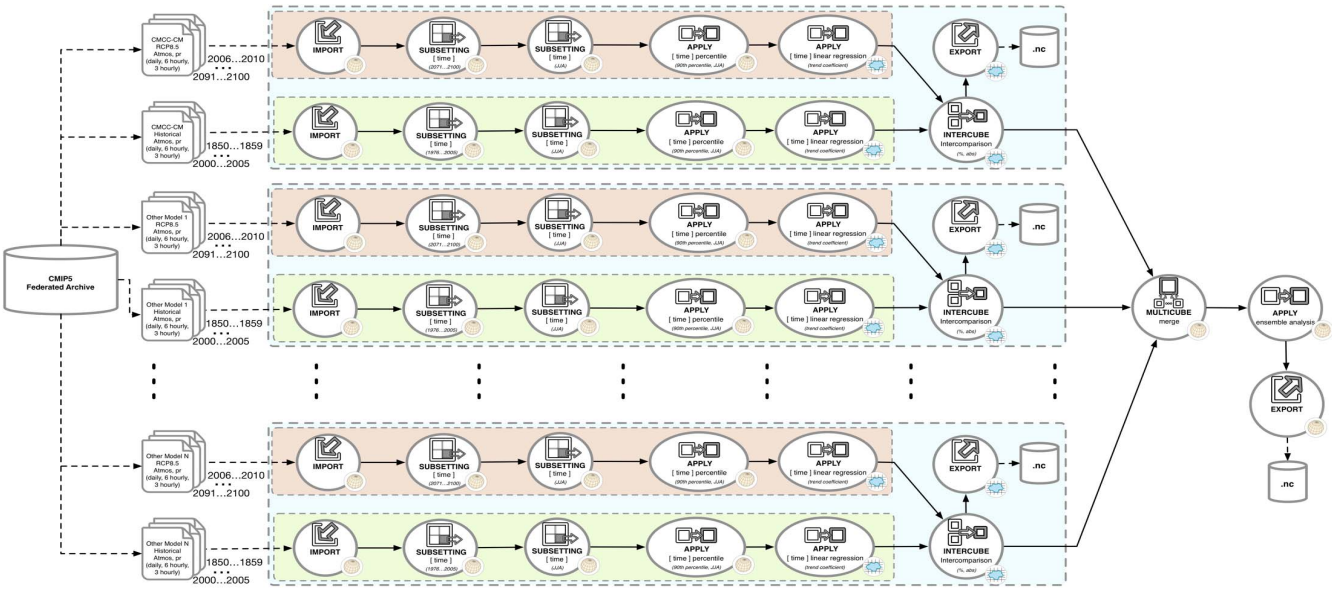

Figure 1. High-level design of the multi-model experiment workflow. The picture highlights the multiple single-model sub-workflows and the final statistical analysis

C. Expected Results

The main research results and impacts associated to this case study are (i) the ability to deal in an easy manner with large scale, massive climate model intercomparison data analysis experiments; (ii) the opportunity to run complex data analysis workflows across multiple data centers, by also integrating well-known existing tools, libraries and command line interfaces; (iii) the possibility to strongly reduce the time-to-solution and complexity associated to this class of large-scale experiments; and (iv) the possibility to address the re-use of final products, intermediate results and workflows.

Finally, it is worth of mentioning that security aspects related to authentication and authorization should be also entirely covered throughout the entire workflow.

D. Use Case on Precipitation Trend Analysis

As a pilot use case for the proposed case study, we have selected the *precipitation trend analysis*, since it is scientifically relevant and it allows validating general infrastructural aspects valid for the other classes of data analysis (e.g. outlier analysis and climate change signal analysis). Fig. 1 shows the workflow to analyse precipitation trend over a given spatial domain by comparing anomalies related to a number of models in the context of CMIP5 Federated Archive.

The experiment consists of two steps:

- in the former there are a number of identical sub-workflows. Each sub-workflow is associated with a specific climate model involved in the CMIP5 experiment and is independent by the others. A future climate scenario must be also defined as input for this step;
- in the latter there is a final workflow performing a statistical analysis on the set of output provided at the end of the first step.

In Fig. 1 the sub-workflows are shown within the bigger rectangles (one for each model). Each of them includes two parallel branches related to the historical and future scenario data processing. Note that time domain related to historical data is fixed. For instance, the range 1976-2005 is adopted for the experiment. Time domain related to models shall have the same duration (e.g. 30 years) though it refers clearly to a future time range like 2071-2100.

The sub-workflows in the first phase of the experiment aim at performing (for each climate model given as input) the following tasks: (i) discovery of the two datasets (historical and future scenario data); (ii) spatio/temporal sub-setting based on the user's input; (iii) evaluation of the precipitation trend for both the datasets separately; (iv) comparison of the trends over the considered domain; and (v) 2D map generation (output).

The multi-model statistical analysis in the second phase of the experiment includes the following four steps: (i) data gathering from the first phase (NetCDF files [15]); (ii) data re-gridding; (iii) statistical analysis; and (iv) final creation of 2D maps related to the inferred statistical indicators.

V. Architectural Design Solution

From an architectural point of view the solution proposed in this work relies on the following layers:

- the *Science Gateway* framework to easily develop, manage and deploy new scientific gateway. For this use case, a specific Data Analytics Science Gateway will be developed.
- the *Workflow Management System* (WfMS) to define, execute, orchestrate and monitor the scientific workflows across multiple sites (coarse-grain level). An additional component will act as the Workflow MarketPlace and will store/publish a set of workflows for the ENES community.
- the *Core Engine* to (i) decouple the Science Gateways from the low-level infrastructure and services and (ii) address multiple types of requests with the same pattern.
- the *Big Data Analytics* component to address large scale analytics needs at the level of a single site (fine-grain level).
- the *Publication Service* component to publish the results of an experiment, making them available to the user community.
- the *Fabric layer* (data index, access and security components) to provide the data search, discovery, access, AuthZ and AuthN needs.

VI. Infrastructural Implementation

From an infrastructural standpoint, the following components have been selected and extended during the INDIGO-DataCloud project lifetime for the implementation of the proposed use case architecture.

As depicted in the Fig. 2, the current implementation consists of the components presented in the following:

(i) the Future Gateway Framework provides the Science Gateway framework to support this case study (a set of portlets have been already implemented in terms of GUI and interaction with the INDIGO-DataCloud PaaS layer);

(ii) Kepler [16] represents a first class solution as WfMS and it has been adopted to run the multi-site workflow for the precipitation trend analysis use case. It also provides through its user interface another GUI to run the experiment;

(iii) the Future Gateway Engine, jointly with JSAGA [17-18] and the related adaptors will act as the Core Engine; the user interface level interacts with this component to both submit the experiment and check its status at run-time. In this regard the two tasks will be performed through specific FutureGateway REST API invocations and dispatched to the involved JSAGA adaptors for back-end management;

(iv) myExperiment [19] has been selected as the proper community-base component to publish and share workflows. It represents the front-end for application-

level, end-users looking for experiments already designed in the community to be re-used as is in other contexts; as back-end, a git repository provides the proper service for the developers community aiming at improving, fixing, extending the workflow document associated to the experiment; the two components represents front-end and back-end of the Marketplace envisioned for this use case;

(v) Ophidia [20] is a big data analytics framework exploited in the infrastructure to run both the first phase of the analytics experiment on the three sites and the multi-model statistical analysis, at the end, on a central site; for performance reasons (e.g. to reduce data movement) each site manages an instance of Ophidia in the same environment of the associated ESGF data node. The big data analytics workflows are executed through a specific JSAGA adaptor from the Future Gateway Engine, which is responsible for submitting the request to the proper Ophidia engine. As an example, Kepler invokes the FutureGateway REST API for the Ophidia sub-workflow submission and monitoring. Ophidia provides a set of parallel operators (based on MPI) a native I/O server running in-memory analytics [21] and exploiting OpenMP. Ophidia supports data analytics workflow management [22]. More details about the high-performance aspects of Ophidia are out of the scope of this paper and can be found in previous works [20-22];

(vi) the OPeNDAP/THREDDS service is being used as publication service;

(vii) the ESGF nodes provide the needed Fabric layer;

(viii) the dynamic instantiation of the services on the testbed sites (e.g. the Ophidia clusters and the Kepler WfMS) is performed through the INDIGO-DataCloud PaaS layer, in particular through the Orchestrator and, on the underneath, the Infrastructure Manager (IM) [23]. The input request for the Orchestrator is a TOSCA-compliant [24] (Topology and Orchestration Specification for Cloud Applications) document defining the setup of a specific software component (e.g. Kepler). Such a document includes references to one or more Ansible [25] roles for the deployment and configuration steps of the associated software components. At the IaaS level, IM supports both native and standard interfaces like OCCI [26] (Open Cloud Computing Interface) for the interaction with the service front-end of the IaaS provider's (OpenNebula, OpenStack or an OCCI-compatible site) internal infrastructure management framework.

It should be noted that, Fig. 2 and Fig. 3 highlight a set of relevant aspects addressed by the proposed solution like (i) the interoperability and integration with ESGF, (ii) the multi-site nature of the experiment, (iii) marketplace for sharing workflows, (iv) the big data frameworks for climate data analysis, and (v) the different lifetime associated to the services (long-running, on-demand, persistent, etc.).

It is worth mentioning that, the use case exploits the INDIGO capabilities in terms of software framework deployed on cloud, as well as the two-level workflow strategy based on Kepler and Ophidia to run geographically distributed, scientific data analysis [27].

In particular:

- the general-purpose Kepler workflow management system is exploited in this use case to orchestrate multi-site tasks (level 1) related to the multi-model part of the experiment;
- the Ophidia framework is adopted at the single-site level to orchestrate the site-specific analytics workflow (level 2), related to the single-model parts of the experiment. Such workflow runs on multiple sites and includes tens of data processing, analysis, and visualization operators in Ophidia, acting at the same time as a single level-1 task in Kepler.

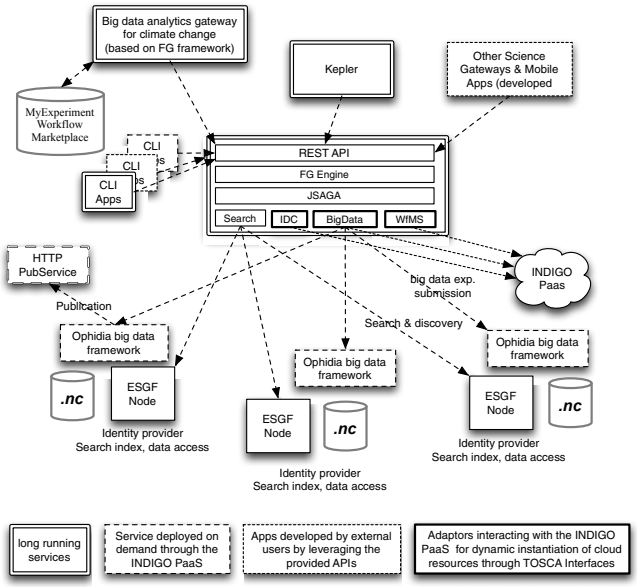

Figure 2. Infrastructural-view of proposed solution.

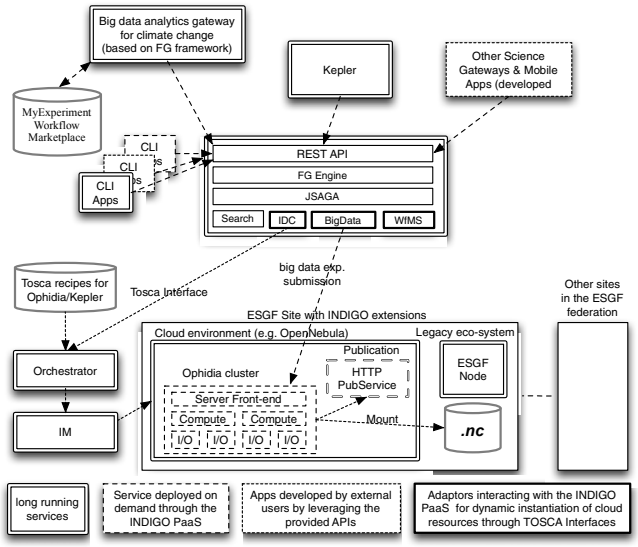

Figure 3. In-depth view of the dynamic instantiation of an Ophidia cluster through the INDIGO PaaS services.

VII. Testbed Setup and Preliminary Insights

A geographically distributed testbed (see Fig. 4) involving a central site (Poznan Supercomputing and Networking Center) for the experiment orchestration and three ESGF sites (Lawrence Livermore National Laboratory, Oak Ridge National Laboratory, and Euro-Mediterranean Center on Climate Change) represents the test environment for the proposed solution that is being applied on CMIP5 datasets.

A set of models has been chosen for the preliminary phase of the testbed by a group of climate scientists involved in the experiment. More specifically the selected models are: CMCC-CM, CMCC-CMS, GISS models, CNRM-CM5, CSIRO-Mk3-6-0, MIROC4h, and GFDL models.

Preliminary runs of the experiment in the testbed demonstrate that *running a multi-model experiment like the one presented in this paper takes order of minutes to be completed, which represents an unprecedented (small) scale with regard to the current state of the art.*

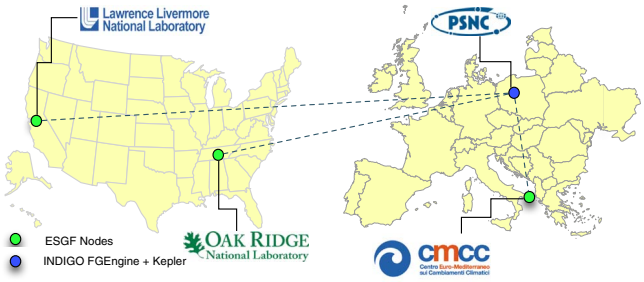

Figure 4. Distributed testbed setup for the precipitation trend analysis use case. Involved sites: LLNL (CA, USA), ORNL (TN, USA), CMCC (IT, EU) as ESGF sites and PSNC (PL, EU) as orchestrator site of the experiment.

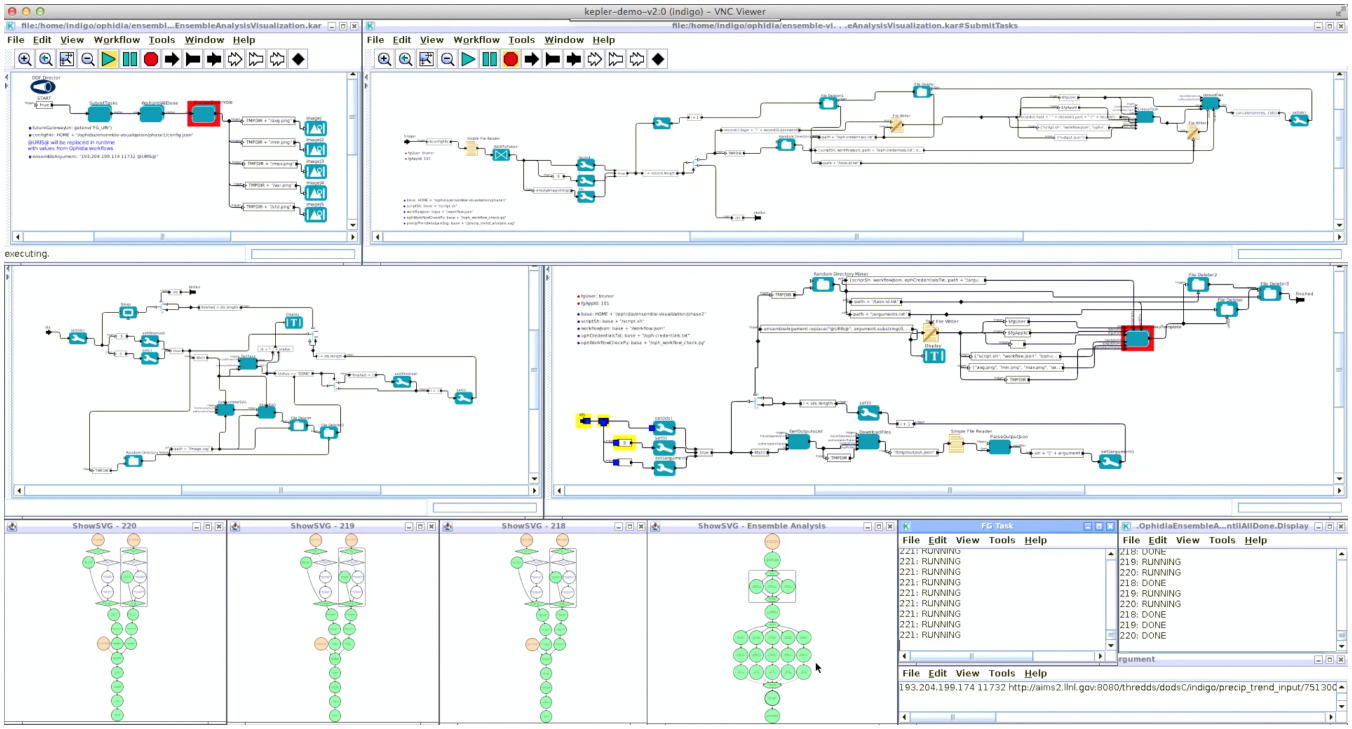

Figure 5. Kepler user interface implemented for the precipitation trend analysis experiment. The interface consists of a set of gadgets that have been designed and setup follwing a dashboard-like approach to provide a real-time monitoring view of the experiment run.

VIII. Kepler User Interface

Fig. 5 shows the Kepler user interface related to the precipitation trend analysis experiment [28]. The interface provides a clear understanding about how the two-level workflow mechanism has been implemented.

A set of gadgets has been designed and setup following a dashboard-like approach. In particular (see Fig. 5) starting from the top left corner:

- the first gadget provides the top-level experiment flow consisting of the following three phases: (phase1) the submission of the analytics sub-workflows to the three ESGF sites, (phase2) the monitoring loop, which acts as a barrier until the end of the first step, and finally (phase3) the statistical analysis on the three output provided by the first step;
- the three following gadgets represent respectively an in-depth monitoring view of the three phases ran by the first gadget;
- at the very bottom, the six remaining gadgets provide respectively: (i) the monitoring of the analytics workflows (at the Ophidia level) running

in parallel on the three ESGF sites during the phase1 of the experiment. The three gadgets displays a fine-grain monitoring view of the analytics tasks and are provided by querying in real-time the different Ophidia instances; (ii) the monitoring of the analytics workflow (at the Ophidia level) running the final statistical analysis (phase3) on a single Ophidia instance, and finally, the last two gadgets provide the coarse-grain status information (SUBMITTED; RUNNING, DONE) from the Kepler perspective about the 4 analytics workflows.

IX. Added Value of the Proposed Solution

The added value of the solution proposed in the INDIGO-DataCloud project are summarized in the following: (i) it implements a different paradigm (from client- to server-side), (ii) it intrinsically reduces data movement, (iii) it makes lightweight the end-user setup, (iv) it fosters re-usability (of data, final/intermediate products, workflows, sessions, etc.) since everything is managed on the server-side, (v) it complements, extends and interoperates with the ESGF stack, (vi) it provides a "tool" for scientists to run multi-model experiments, and finally, (vii) it can drastically reduce the time-to-solution for these experiments from weeks to hours.

At the time the paper is being written, *the proposed testbed represents the first concrete implementation of a distributed multi-model experiment in the ESGF/CMIP context joining server-side and parallel processing, end-to-end workflow management and cloud computing.*

X. Conclusions and Future Work

This paper presents a case study on *climate models intercomparison data analysis* implemented in the context of the EU H2020 INDIGO-DataCloud project.

It addresses scientific challenges associated to multiple classes of data analysis like trend analysis (specifically targeted in this paper), anomalies analysis, and climate change signal analysis.

As opposed to the current scenario based on search & discovery, data download, and client-based data analysis, the INDIGO-DataCloud architectural solution described in this paper addresses the scientific requirements discussed in Section II by providing a paradigm shift based on server-side and high performance big data frameworks jointly with two-level workflow management systems realized at the PaaS level via a Cloud infrastructure; it joins at the architectural level *cloud computing*, *HPC* and *big data*.

As discussed in this paper the proposed approach allows overcoming current limitations regarding client & sequential data analysis, static setup approaches, poor performance as well as a complete lack of workflow support, and domain-oriented big data approaches to enable large scale, high performance multi-model climate data analysis experiments.

This work is being tested on CMIP5 data, but it significantly contributes to an analytics-aware ESGF infrastructure for CMIP6.

CMIP6 preparatory activity is still ongoing; the estimated data volume in CMIP6 will be around 20-30 times bigger than CMIP5, so it will represent a strong test case for the proposed solution. The requirement of elasticity, as the quick provision and release of additional resources according to the workload, as well as the reconfiguration of the virtual infrastructure will even be more relevant. The CMIP6 implementation phase will be closely followed in the next months to further apply, test and validate the approach and the solution presented in this work.

As a concluding remark, the solution proposed in this paper aims at providing a core infrastructural piece still missing in the current climate scientists' research eco-system.

Future work will mainly relate to a larger exploitation of the proposed solution across ESGF sites and to the implementation of a dedicated Science Gateway as a central hub for scientists to run multi-model climate data analysis experiments.

Acknowledgment

This work was supported by the EU H2020 INDIGO-DataCloud Project (Grant Agreement 653549), by the EU H2020 EUBra-BIGSEA Project (Grant Agreement 690116), and by the EU FP7 IS-ENES2 Project (Grant Agreement 312979). This work was also supported by the U.S. Department of Energy, Office of Science, under Contract DE-AC02-06CH11357 and in part by the resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725; and also in part by the Accelerated Climate Modeling for Energy (ACME) program, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research.

References

[1] J. Dongarra, P. Beckman, et al.: "The International Exascale Software Project roadmap". International Journal of High Performance Computing Applications (IJHPCA) 25(1): 3-60 (2011), ISSN 1094-3420, doi: 10.1177/1094342010391989.

[2] European Exascale Software Initiative roadmap - http://www.eesi-project.eu/pages/menu/project/eesi-1/publications/final-report-recommendations-roadmap.php

[3] PRACE – The Scientific Case for High Performance Computing in Europe 2012-2020 - http://www.prace-ri.eu/IMG/pdf/prace_-_the_scientific_case_-_full_text_-.pdf

[4] Earth System Grid Federation - http://esgf.llnl.gov

[5] L. Cinquini, *et al.*: "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data". Future Generation Comp. Syst. 36: 400-417 (2014)

[6] ENES European Network for Earth System Modelling

[7] Intergovernmental Panel on Climate Change – http://www.ipcc.ch

[8] IPCC Fifth Assessment Report - https://www.ipcc.ch/report/ar5/

[9] Working Group on Coupled Modelling - http://www.wcrp-climate.org/wgcm/

[10] World Climate Research Programme - http://www.wcrp-climate.org/

[11] INDIGO-DataCloud - https://www.indigo-datacloud.eu

[12] D. Salomoni, I. Campos Plasencia, *et. al*: "INDIGO-Datacloud: foundations and architectural description of a Platform as a Service oriented to scientific computing". CoRR abs/1603.09536 (2016)

[13] Research Data Alliance (RDA) - www.rd-alliance.org

[14] "CONFIRMATION OF SUPPORT TO INITIAL REQUIREMENTS AND EXTENDED LIST - D2.4", INDIGO-DataCloud EU Deliverable D2.4, INDIGO-WP2-D2.4-V20 - https://owncloud.indigo-datacloud.eu/index.php/s/pLCB3uSwlmgtjH

[15] Rew, R. K. and G. P. Davis, "The Unidata netCDF: Software for Scientific Data Access", 6th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, Anaheim, California, American Meteorology Society, pp. 33-40, February 1990.

[16] M. Płóciennik, T. Żok, et al. 2013. "Approaches to Distributed Execution of Scientific Workflows in Kepler". Fundam. Inf. 128, 3 (July 2013), 281-302. DOI=http://dx.doi.org/10.3233/FI-2013-947.

[17] The JSAGA website. http://grid.in2p3.fr/jsaga. Accessed: February, 04. 2016.

[18] The SAGA OGF Standard Specification. http://www.ogf.org/documents/GFD.90.pdf.

[19] De Roure, D., Goble, C. and Stevens, R. (2009): "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows". Future Generation Computer Systems25, pp. 561-567. doi:10.1016/j.future.2008.06.010.

[20] S. Fiore, A. D'Anca, C. Palazzo, I. T. Foster, D. N. Williams, G. Aloisio: "Ophidia: Toward Big Data Analytics for eScience". ICCS 2013: 2376-2385.

[21] D. Elia, *et al.*. 2016. "An in-memory based framework for scientific data analytics". In Proceedings of the ACM International Conference on Computing Frontiers (CF '16). ACM, New York, NY, USA, 424-429. DOI: http://dx.doi.org/10.1145/2903150.2911719.

[22] C. Palazzo, *et al.*: "A workflow-enabled big data analytics software stack for escience". HPCS 2015: 545-552.

[23] M. Caballer, I. Blanquer, G. Molto, and C. de Alfonso, "Dynamic Management of Virtual Infrastructures." Journal of Grid Computing 13(1): 53–70 (2015) - http://link.springer.com/article/10.1007/s10723-014-9296-5.

[24] TOSCA http://docs.oasis-open.org/tosca/TOSCA/v1.0/os/TOSCA-v1.0-os.html.

[25] Ansible - https://www.ansible.com/

[26] Open Cloud Computing Interface http://occi-wg.org/about/specification/

[27] M. Płóciennik, S. Fiore, *et al.*: "Two-level Dynamic Workflow Orchestration in the INDIGO DataCloud for Large-scale, Climate Change Data Analytics Experiments". ICCS 2016: 722-733.

[28] Multi-model, distributed analytics experiment in INDIGO. https://www.youtube.com/watch?v=Xqr3JRc_B10&feature=youtu.be